

# Appendix

1		
2	Appendix A Related Work	1
3	A.1 Trigger Design . . . . .	1
4	A.2 Sample Selection . . . . .	1
5	Appendix B Preliminaries	2
6	B.1 Model Training . . . . .	2
7	B.2 Poison-only Clean-label Backdoor Attacks . . . . .	2
8	B.2.1 Attack Knowledge . . . . .	2
9	B.2.2 Attack Workflow . . . . .	2
10	Appendix C Algorithms with other negative functions	3
11	Appendix D Gradient Magnitude Similarity Deviation	4
12	Appendix E Human Visual System	5
13	E.1 Distinct Sensitivity to RGB . . . . .	5
14	Appendix F The Effect of Triggers on Backdoor Attacks	7
15	Appendix G Details of Experiment Setting	9
16	Appendix H Extended Ablation Study	10
17	H.1 Effect of Target Label . . . . .	10
18	H.2 Category Similarity . . . . .	11
19	H.3 Applying our methods to poisoned-label backdoor attacks . . . . .	13
20	H.4 The effect of Component A in Tiny-Imagenet . . . . .	14
21	Appendix I Stealthiness of our components on multiple attacks	15

## 22 A Related Work

23 In backdoor attacks, the adversary aims to embed a designed trigger in the victim model. Therefore,  
 24 the poisoned models misclassify the trigger-embedded samples to the predefined target label (Gu  
 25 et al. [2017], Chen et al. [2017]) while maintaining high accuracy for unaltered inputs. Multiple  
 26 backdoor attacks prove their effectiveness in multimodal learning (Wang et al. [2024], Han et al.  
 27 [2024]), federated learning (Li et al. [2023], Chen et al. [2023]), diffusion model (Chou et al. [2023],  
 28 Li et al. [2024]), dataset distillation (Liu et al. [2023]), and other scenarios (Zhao et al. [2024]).

29 Among current backdoor attacks, Poison-only Backdoor Attacks (PBAs) have attracted huge attention  
 30 given their widespread use and ease of construction in real-world scenarios (Li et al. [2021], Qi  
 31 et al. [2023]). PBAs poison the models by merely manipulating the training dataset, in which the  
 32 effectiveness of attacks hinges on Trigger Design and Sample Selection.

### 33 A.1 Trigger Design

34 The detectable of simple-designed visible triggers in traditional attacks (Gu et al. [2017], Chen  
 35 et al. [2017]) by both humans and machines leads the adversary to focus on the design of invisible  
 36 triggers and physical triggers. In computer vision (CV), invisible triggers involve incorporating minor  
 37 perturbations by tweaking the pixel values and positions of the original image (Bai et al. [2022]).  
 38 Despite its stealthiness, the constraint of invisibility poses a significant limitation and conflict in ASR.  
 39 Therefore, Wenger et al. [2022] introduces natural triggers based on the hypothesis that there may be  
 40 naturally occurring physically colocated objects already present in popular datasets such as ImageNet.  
 41 Furthermore, Lin et al. [2020] proposes a trigger formulated from a combination of existing benign  
 42 features to bypass the machine detection.

43 Efforts to overcome the dilemma frequently result in unsatisfactory performance (e.g., high injection  
 44 rates, ineffective backdoor embeddings, limited transferability, and weakened robustness). For in-  
 45 stance, Wang et al. [2022] introduces BppAttack, a stealthy attack that leverages image quantization  
 46 and dithering to induce triggers. Given the constrained effectiveness of imperceptible modifications,  
 47 adversaries struggle to enhance the ASR by employing adversarial training combined with label  
 48 flipping. Recently, (Gao et al. [2024]) formulates a bi-level optimization problem to balance the  
 49 conflict of ASR and stealthiness with sparsity and invisibility constraints. The upper-level optimiza-  
 50 tion problem aims to minimize the loss on poisoned samples by optimizing the trigger. Meanwhile,  
 51 the lower-level problem focuses on minimizing the loss across all training samples through the  
 52 optimization of model weights, which deviates from a poison-only attack.

53 **Summary** Current PBAs primarily focus on the design of triggers, leading to multiple triggers  
 54 that exhibit unique advantages under different metrics (e.g., design complexity, feature intensity, the  
 55 ability to bypass defenses, stealthiness, and dataset dependency). Therefore, **it is valuable to explore**  
 56 **generalization optimization strategies to enhance various triggers on both ASR and stealthiness.**  
 57 Additionally, **current research overlooks the effect of sample selection in the design process.**

### 58 A.2 Sample Selection

59 Clean-label backdoor attacks are seen as the stealthiest attacks, as adversaries can only poison samples  
 60 from the target class without changing their labels. The dilemma of unsatisfactory ASR of current  
 61 PBAs that merely depend on the trigger design led to the research study of sample selection. Gao  
 62 et al. [2023] reveals differential sample importance and selects “hard” samples via three metrics (e.g.,  
 63 Forgetting Event (as depicted in **Section 2.1**), Loss Value, and Gradient Norm) to enhance the PBAs.  
 64 The poisoned models tend to learn the implicit projection between the trigger feature and the target  
 65 label to evade the difficulty of the original classification upon such “hard” samples. Details of Loss  
 66 Value and Gradient Norm can be seen as follows.

67 **Loss Value** Given a benign model  $f_\theta$  (trained on the benign training set  $D_{tr}$ ), the loss value of  
 68 model on sample  $(x_i, y_i)$  can be represented as  $L(f_\theta(x_i), y_i)$ . We choose samples with the greatest  
 69  $\alpha * |D_{tr}|$  values in the subset  $D_t$  are chosen for poisoning:

$$D_s = \arg \max_{D_s \subset D_t} \sum_{(x_i, y_i) \in D_s} L(f_\theta(x_i), y_i). \quad (1)$$

70 **Gradient Norm** Given a benign model  $f_\theta$  (trained on the benign training set  $D_{tr}$ ), the  $l_2$ - gradient  
 71 norm of model on sample  $(x_i, y_i)$  can be represented as  $\|\nabla_\theta L(f_\theta(x_i), y_i)\|_2$ . We choose samples  
 72 with the greatest  $\alpha * |D_{tr}|$  values in the subset  $D_t$  are chosen for poisoning:

$$D_s = \arg \max_{D_s \subset D_t} \sum_{(x_i, y_i) \in D_s} \|\nabla_\theta L(f_\theta(x_i), y_i)\|_2. \quad (2)$$

73 Han et al. [2024] further improves the efficiency of attacks based on an optimized backdoor gradient-  
 74 based score. Moreover, Hayase and Oh [2022] formulates sample selection as a bi-level optimization  
 75 problem: construct strong poison examples that maximize the ASR. Furthermore, some scientists  
 76 propose novel sample selection methods based on poisoning masks (Zhu et al. [2023]), confidence-  
 77 based scoring (Wu et al. [2023]), and high-frequency energy (Xun et al. [2024]).

78 **Summary** Current research on sample selection focuses on designing new metrics or training  
 79 derivations to construct data-efficiency attacks, **overlooking the synergistic effect between triggers**  
 80 **and sample selection on ASR enhancement**. Meanwhile, **current methods overlook the effect of**  
 81 **sample selection on stealthiness enhancement**.

## 82 B Preliminaries

### 83 B.1 Model Training

84 The model output function of the image classification can be denoted by  $f_\theta : X \rightarrow Y$ , where  
 85  $x \in X = \{0, 1, \dots, 255\}^{C \times H \times W}$  represents an image domain,  $Y = \{y_1, y_2, \dots, y_k\}$  is a set  
 86 of  $k$  classes, and  $\theta$  denotes the parameters that a DNN learned from the begin training dataset  
 87  $D_{tr} = \{(x_i, y_i)\}_{i=1}^N$ . The benign training with  $D_{tr}$  can be seen as a single-level optimization  
 88 problem. The optimization seeks a model  $f_\theta$  by solving the following problem during training:

$$\min_{\theta} L(D_{tr}, f_\theta) = \sum_{i=1}^{N_{tr}} l(x_i, y_i, f_\theta), \quad (3)$$

89 where  $l$  is the loss function (e.g., the cross-entropy), and  $(x_i, y_i) \in D_{tr}$ .

### 90 B.2 Poison-only Clean-label Backdoor Attacks

#### 91 B.2.1 Attack Knowledge

92 In a poison-only backdoor attack, an adversary has access to the original training dataset  $D_{tr}$  and is  
 93 allowed to inject the pre-defined trigger into a small subset of the training set. Specifically, attacks  
 94 can be called clean-label attacks if the adversary does not change the ground-truth label of the  
 95 origin data. Furthermore, the adversary has no knowledge and the ability to modify other training  
 96 components (e.g., loss functions, model architecture, training schedule, optimization algorithm,  
 97 etc). Consequently, attackers can only influence model weights through data poisoning. The latent  
 98 connection between the trigger and the target label is learned only during the training process. In  
 99 the inference stage, we assume that the adversary lack access to the prediction vectors. In general,  
 100 poison-only clean-label attacks require minimal capacities and therefore can be applied in many  
 101 real-world scenarios.

#### 102 B.2.2 Attack Workflow

103 We detail the workflow of poison-only clean-label backdoor attacks to formalize the theoretical  
 104 foundations. How to generate the poisoned dataset  $D_p$  is the cornerstone of the attack. Details about  
 105 the attack knowledge of poison-only clean-label backdoor attacks can be seen at Appendix B. We  
 106 remark on the important evaluation criteria at the following steps.

107 **Step 1: Select samples to be poisoned (by attackers).**  $D_p$  consists of two disjoint parts. Given a  
 108 target label  $y_t$ , a subset  $D_s$  is selected from target-label set  $D_t = \{(x_i, y_i) | (x_i, y_i) \in D_{tr}, y_i = y_t\}$  to  
 109 be poisoned and the remain benign samples can be denoted as  $D_b = D_{tr} \setminus D_s$ . Here we define a binary  
 110 vector  $M = [M_1, M_2, \dots, M_{|D_{tr}|}] \in \{0, 1\}^{|D|}$  to represent the poisoning selection. Specifically,

111  $M_i = 1$  indicates that  $x_i$  is selected to be poisoned while  $M_i = 0$  means the benign sample. We  
 112 denote  $\alpha := \frac{|D_s|}{|D_{tr}|}$  as the poisoning rate. Note that most existing backdoor attack methods randomly  
 113 select  $\alpha \cdot |D_{tr}|$  samples to be poisoned.  $\alpha$  serves as a crucial indicator of stealthiness in poison-only  
 114 attacks. Backdoor attacks are supposed to maintain a high attack success rate with  $\alpha$  as small as  
 115 possible to evade both machine and manual inspections.

116 **Step 2: Trigger Insertion (by attackers).** In computer vision applications, the adversary designs  
 117 a trigger pattern  $w$  by tweaking the pixel values and positions of the benign image. The generator  
 118 of poisoned images can be denoted as  $f_g : X \rightarrow X$ . For example,  $f_g(x) = (1 - m) * x + m * w$ ,  
 119 where the mask  $m \in [0, 1]^{C \times H \times W}$  representing the poison area of the trigger  $w$  and  $*$  representing  
 120 the element-wise product. Therefore, given the target label  $y_t$  in a clean-label attack, the generated  
 121 poisoned training dataset could be denoted as  $D_p = \{(x_i, y_i) |_{if\ m_i=0}, \text{ or } (f_g(x_i), y_t) |_{if\ m_i=1}\}_{i=1}^{|D_{tr}|}$ .  
 122 For stronger stealthiness, the trigger  $w$  is expected to be sufficiently invisible, which means the  
 123 distance  $L_D(f_g(x_i), x_i)$  should be small.

124 **Step 3: Model Training (by users).** Once the poisoned dataset  $D_p$  is generated, users will train the  
 125 poisoned DNN via the period described in section 3.1.1. The stealthiness and utility of backdoor  
 126 attacks demand imperceptible dataset modifications, requiring the poisoned model  $\tilde{f}_\theta$  to maintain  
 127 high accuracy on benign test data. Otherwise, users would not adopt the poisoned model and no  
 128 backdoor could be implanted. The accuracy on clean test set  $D_{clean}$  can be computed by:

$$CleanACC = \frac{1}{N_{clean}} \sum_{i=1}^{N_{clean}} ACC(\tilde{f}_\theta(x_i), y_i) \quad (4)$$

129 where  $N_{clean}$  means the number of clean test set.  $(x_i, y_i) \in D_{clean}$  and  $y_i$  is the ground-truth label.  
 130  $ACC(y_{pre}, y)$  will be set to 1 if  $y_{pre} = y$  and 0 otherwise.

131 **Step 4: Activate the backdoor using the trigger during the inference stage (by attackers).** The  
 132 attackers expect to activate the injected backdoor using the trigger  $w$  defined in step 2. Given the  
 133 poisoned model  $\tilde{f}_\theta$ , the Attack Success Rate (ASR) of a backdoor attack can be computed by:

$$ASR = \frac{1}{N_{clean}} \sum_{i=1}^{N_{clean}} ACC(\tilde{f}_\theta(f_g(x_i)), y_t) \quad (5)$$

134 where  $N_{clean}$  means the number of clean test set  $D_{clean}$ .  $f_g(x_i)$  represents the poisoned image on  
 135 image  $x_i$  and  $y_t$  is the target label.  $\tilde{f}_\theta$  and  $ACC(y_{pre}, y)$  are defined in Step 3.

## 136 C Algorithms with other negative functions

137 **Algorithm 1** Metric Calculation with Negative Function  $N_F$  at  $O(n)$

---

138 **Input :** Train Dataset  $D_{tr}$ , Target Label  $y_t$ , Misclassification Events  $N_e((x_i, y_i), y_m)$   
 139 **Output :** Calculated Metric of Samples  
 140 **for** image  $(x_i, y_t) \in D_{tr}$  **do**  
 141      $Num[y_m], Sum = 0$   
 142     **for**  $y_m \in Y$  **do**  
 143          $Num[y_m] = Num[y_m] + N_e((x_i, y_t), y_m)$   
 144          $Sum = Sum + Num[y_m]$   
 145     **end for**  
 146     **end for**  
 147     **for**  $y_m \in Y$  **do**  
 148          $Cls[y_m] = 1 - \frac{Num[y_m]}{Sum}$   
 149     **end for**  
 150     **for** image  $(x_i, y_t) \in D_{tr}$  **do**  
 151          $Metric[x_i] = 0$   
 152         **for**  $y_m \in Y$  **do**  
 153              $Metric[x_i] = Metric[x_i] + Cls[y_m] * N_e((x_i, y_t), y_m)$   
 154         **end for**  
 155     **end for**

---



---

```

156 Algorithm 2 Metric Calculation with Negative Function  $N_F$  at  $O(n^2)$ 
157 Input : Train Dataset  $D_{tr}$ , Target Label  $y_t$ , Misclassification Events  $N_e((x_i, y_i), y_m)$ 
158 Output : Calculated Metric of Samples
159 for image  $(x_i, y_t) \in D_{tr}$  do
160    $Num[y_m] = 0$ 
161   for  $y_m \in Y$  do
162      $Num[y_m] = Num[y_m] + N_e((x_i, y_t), y_m)$ 
163   end for
164 end for
165 for  $y_m \in Y$  do
166    $Sum = Sum + Num[y_m] * Num[y_m]$ 
167 end for
168 for  $y_m \in Y$  do
169    $Cls[y_m] = 1 - \frac{Num[y_m] * Num[y_m]}{Sum}$ 
170 end for
171 for image  $(x_i, y_t) \in D_{tr}$  do
172    $Metric[x_i] = 0$ 
173   for  $y_m \in Y$  do
174      $Metric[x_i] = Metric[x_i] + Cls[y_m] * N_e((x_i, y_t), y_m)$ 
175   end for
176 end for

```

---

```

177 Algorithm 3 Metric Calculation with Negative Function  $N_F$  at  $O(e^n)$ 
178 Input : Train Dataset  $D_{tr}$ , Target Label  $y_t$ , Misclassification Events  $N_e((x_i, y_i), y_m)$ 
179 Output : Calculated Metric of Samples
180 for image  $(x_i, y_t) \in D_{tr}$  do
181    $Num[y_m] = 0$ 
182   for  $y_m \in Y$  do
183      $Num[y_m] = Num[y_m] + N_e((x_i, y_t), y_m)$ 
184   end for
185 end for
186 for  $y_m \in Y$  do
187    $Sum = Sum + exp(-Num[y_m])$ 
188 end for
189 for  $y_m \in Y$  do
190    $Cls[y_m] = 1 - \frac{exp(-Num[y_m])}{Sum}$ 
191 end for
192 for image  $(x_i, y_t) \in D_{tr}$  do
193    $Metric[x_i] = 0$ 
194   for  $y_m \in Y$  do
195      $Metric[x_i] = Metric[x_i] + Cls[y_m] * N_e((x_i, y_t), y_m)$ 
196   end for
197 end for

```

---

## 198 D Gradient Magnitude Similarity Deviation

199 Images visually insensitive to triggers are selected by calculating the GMSD between benign images  
200 and poisoned images to conceal the trigger feature in the target-label feature. GMSD is a full-reference  
201 image quality assessment (FR-IQA) model that leverages pixel-wise gradient magnitude similarity  
202 (GMS) to quantify local image quality and the standard deviation of the global GMS map to quantify  
203 the final image quality. Specifically, the gradient magnitude is derived using the Prewitt filter, which  
204 estimates horizontal  $x$  and vertical  $y$  gradient components via convolution by the following kernels:

$$h_x = \begin{bmatrix} 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \end{bmatrix}, \quad h_y = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 \\ -1/3 & -1/3 & -1/3 \end{bmatrix} \quad (6)$$

205 Convolving  $h_x$  and  $h_y$  with the reference and distorted images yields the horizontal and vertical  
 206 gradient images of  $r$  and  $d$ .  $m_r(i)$  and  $m_d(i)$  represent the gradient magnitudes of  $r$  and  $d$  at location  
 207  $i$ , which can be computed as follows:

$$m_r(i) = \sqrt{(r \otimes h_x)^2(i) + r \otimes h_y)^2(i)}, \quad m_d(i) = \sqrt{(d \otimes h_x)^2(i) + d \otimes h_y)^2(i)} \quad (7)$$

208 where symbol " $\otimes$ " denotes the convolution operation. The gradient magnitude similarity (GMS) map  
 209 is computed based on the gradient magnitude images  $m_r(i)$  and  $m_d(i)$  as follows:

$$GMS(i) = \frac{2m_r(i)m_d(i) + c}{m_r^2(i) + m_d^2(i) + c} \quad (8)$$

210 where  $c$  is a positive constant that supplies numerical stability. Gradient Magnitude Similarity Mean  
 211 (GMSM) serves as the local quality map (LQM) of the distorted image  $d$  with average pooling applied  
 212 to assume that each pixel has the same importance in estimating the overall image quality:

$$GMSM = \frac{1}{N} \sum_{i=1}^N GMS(i) \quad (9)$$

213 where  $N$  is the total number of pixels in the image. Clearly, a higher GMSM score means higher  
 214 image quality. Based on the idea that the global variation of image local quality degradation can  
 215 reflect its overall quality, Gradient Magnitude Similarity Deviation (GMSD) is proposed to compute  
 216 the standard deviation of the GMS map as the final IQA index:

$$GMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N (GMS(i) - GMSM)^2} \quad (10)$$

217 GMSD serves as a quantitative measure of the spatial distribution of distortion severity within an  
 218 image. Specifically, higher GMSD values indicate a wider range of distortion magnitudes across  
 219 local regions, which correlates with degraded perceptual quality due to the exacerbated spatial  
 220 inconsistency of degradation effects.

## 221 **E Human Visual System**

222 Computers encode image colors based on the three primary color channels (RGB). However, current  
 223 design of triggers neglects the differences in human visual perception (Land and McCann [1971]) and  
 224 machine representation. Therefore, knowledge of the human visual system (HVS) can assist adversary  
 225 in more scientifically leveraging the disparities between the human eye and machine systems to  
 226 enhance the stealthiness and functionality of triggers in backdoor attacks.

### 227 **E.1 Distinct Sensitivity to RGB**

228 The human retina contains three types of cone cells, each playing a crucial role in color vision by  
 229 being sensitive to different wavelengths of light. These three types of cone cells work together to  
 230 provide us with color vision. Each type of cone cell contains a different photopigment that is sensitive  
 231 to a specific range of wavelengths. When light enters the eye and stimulates these cone cells, they  
 232 send signals to the brain, which then processes this information to produce our perception of color.

#### 233 **Long-Wavelength Sensitive (L) Cone Cells:**

- 234 • These cone cells are most responsive to long-wavelength light, with a peak sensitivity around  
 235 560 nm, which corresponds to the yellow-green region of the visible spectrum.
- 236 • They are often referred to as "red" cone cells because of their relative sensitivity to longer  
 237 wavelengths, although their peak is not precisely at the red end of the spectrum.
- 238 • L cone cells are abundant in the retina and are essential for distinguishing between colors in  
 239 the red-yellow-green range.

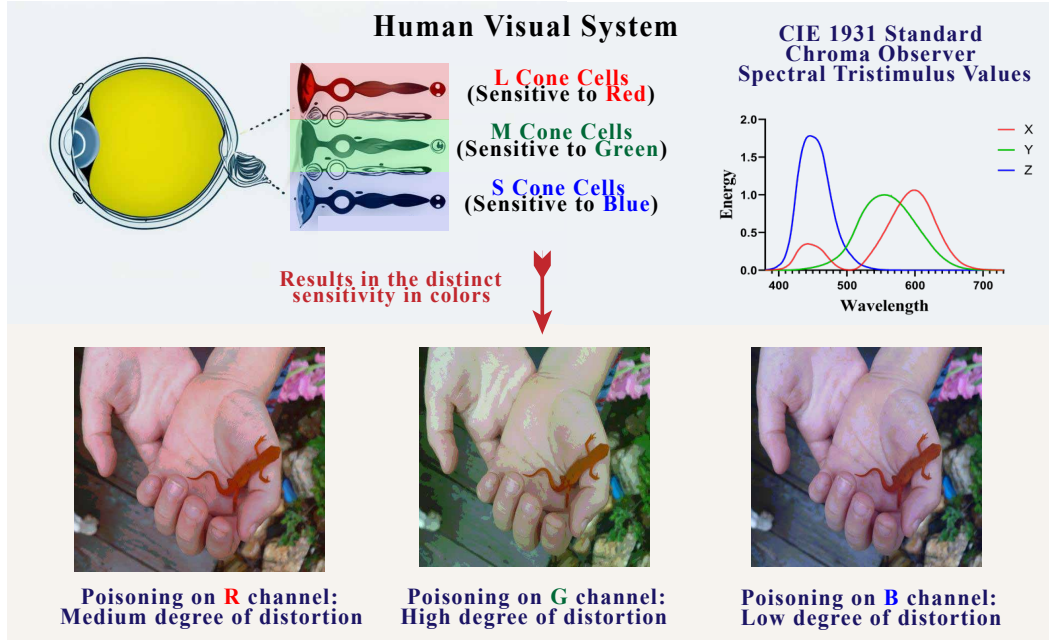


Figure 1: Distinct Sensitivity to Colors in Human Visual System.

#### Medium-Wavelength Sensitive (M) Cone Cells:

- M cone cells have their peak sensitivity around 530 nm, in the green region of the spectrum.
- These cone cells are crucial for perceiving colors in the green range and are involved in color discrimination tasks that require distinguishing between different shades of green and yellow.
- Together with L cone cells, M cone cells form the basis for our perception of a wide range of colors in the visible spectrum.

#### Short-Wavelength Sensitive (S) Cone Cells:

- S cone cells are most responsive to short-wavelength light, with a peak sensitivity around 420 nm, which corresponds to the blue-violet region of the spectrum.
- They are often referred to as "blue" cone cells and are essential for perceiving colors in the blue range.
- S cone cells are less abundant in the retina compared to L and M cone cells, but they play a critical role in our ability to distinguish between colors that have a blue component.

The RGB color system is based on the three primary colors of human vision. Experiments have revealed that when certain spectral colors are represented using the color-matching functions of the RGB color system, negative values emerge. This implies that there are spectral colors that cannot be expressed using the visual primary colors RGB. Therefore, the XYZ color space system in the International Commission on illumination (CIE-XYZ) is introduced to address the dilemma.

We use  $\{R, G, B\}$  to represent value of pixels in the three color channels  $\{x^R, x^G, x^B\}$ . The core objective of the CIE-RGB system is to establish an anchored relationship between color and physical parameters, ensuring a one-to-one correspondence between color perception and tristimulus values. Its design focuses on color appearance through the proportioning of the three primary colors, rather than directly quantifying the sensitivity of the human visual system. The phenomenon that human eyes are most sensitive to green light (555nm) is reflected in the subsequent CIE-XYZ system through the luminance function  $f_Y = 0.2126R + 0.7152G + 0.0722B$ , but this weight distribution is a characteristic of the CIE-XYZ system, not the original design of the CIE-RGB system.

In 1931, CIE standardized conversion relationships between the two systems to resolve the RGB system’s negative value issue, guaranteeing positive tristimulus values in XYZ. Converting RGB values to CIE-XYZ tristimulus values follows a standardized process and the overall process of selecting samples can be outlined step-by-step below:

**Step 1: Normalize CIE-RGB values.** Step 1 aims to convert the value of image  $(R, G, B)$  to the range  $[0, 1]$  :

$$x_{norm}^c = \frac{x^c}{R + G + B}, c \in \{R, G, B\} \quad (11)$$

Specifically, we use  $\{r, g, b\}$  to represent the normalized result  $\{x_{norm}^R, x_{norm}^G, x_{norm}^B\}$ .

**Step 2: Convert normalized CIE-RGB to normalized CIE-XYZ.** The conversion formulas of chromaticity coordinate conversion can be denoted as:

$$\begin{cases} X = (0.490r + 0.310g + 0.200b) / (0.607r + 1.132g + 1.200b) \\ Y = (0.117r + 0.812g + 0.010b) / (0.607r + 1.132g + 1.200b) \\ Z = (0.000r + 0.010g + 0.990b) / (0.607r + 1.132g + 1.200b) \end{cases} \quad (12)$$

CIE 1931 Standard Chroma Observer Spectral tristimulus Values, abbreviated as CIE Standard Chroma Observer, characterizes human ocular spectral sensitivity across wavelengths, as depicted in Figure 1. Furthermore, humans exhibit limited sensitivity to blue light because the blue-sensitive cone cells comprise merely 5% in the human visual system.

**Summary** Based on the above observations, it is appropriate to reassign the poisoning intensity of the trigger design with a particular enhanced poisoning intensity in the blue channel.

## F The Effect of Triggers on Backdoor Attacks

Table 1: Performance of global-poisoning attacks by poisoning 2.5% samples of CIFAR-10.

Attack			Metric		Attack Setting		
Type	no.	Method	ASR	BA	Clean-label	Training Control	Stealthy
Benchmark	a	Benign	-	<b>95.0</b>	✓	✗	✓
	b	Base	<b>8.2</b>	94.8	✓	✗	✓
	c	BppAttack	<b>12.5</b>	94.5	✗	✓	✓
	d	Blended-C	<b>66.4</b>	94.3	✓	✗	✗
MultiBpp (our methods)	e	255:255:8	68.6	94.8	✓	✗	✓
	f	255:255:12	60.0	<b>94.9</b>	✓	✗	✓
	g	24:48:8	<b>76.6</b>	94.7	✓	✗	✓
	h	36:72:12	57.7	94.6	✓	✗	✓
MultiBpp (others)	i	8:255:255	<b>84.1</b>	<b>94.7</b>	✓	✗	✗
	j	255:8:255	72.2	94.3	✓	✗	✗
	k	12:255:255	67.6	94.5	✓	✗	✗
	l	255:12:255	73.8	94.5	✓	✗	✗

**The effect of triggers in MultiBpp attacks** According to Table 1, the red and green channels demonstrate superior attack performance with 84.1% ASR by poisoning at the red channel and 72.2% ASR by poisoning at the green channel. The differential learning sensitivities imply that the model can infer that the feature in the red and green channels are more valuable. Tables 1 {e,f} and {g,h} indicate that increasing the quantization step improves ASR. However, MultiBpp with the quantization intensity of (36 : 72 : 12) yields a lower ASR of 57.7%, compared to 60.0% achieved with 255 : 255 : 12. We hypothesize that the learning effectiveness of the poisoning feature is not solely influenced by the quantization step. Specifically, in scenarios like {f,h}, the model needs to focus on features from all three channels when learning under the configuration of h, whereas it only needs to attend to feature from one channel when learning the trigger feature.

As depicted in Figure 2, the original BppAttack randomly selects data for poisoning. To maintain the stealthiness of the trigger, BppAttack needs to adopt a smaller quantization step (32 : 32 : 32), which makes it difficult for the trigger feature to be learned. We optimize the BppAttack based on



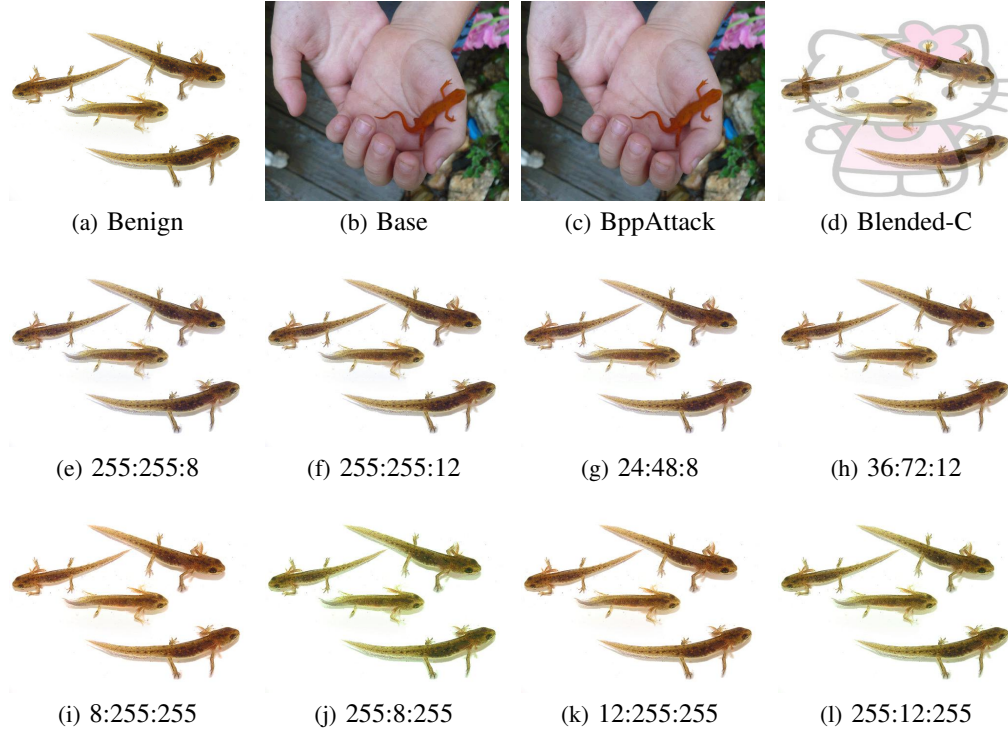


Figure 2: Visualizations of images in global-poisoning attacks. Compared to the benchmark (the first line), images that are visually insensitive to MultiBpp are selected in Component B.  $N_R : N_G : N_B$  represent the distinct quantization intensity in  $R : G : B$  channels.

two key observations. Firstly, current research on colorimetry reveals that the human visual system exhibits vastly different sensitivities to colors. For example, we can observe that enhanced attacks by increasing the poisoning intensity in the blue channel can still maintain invisibility to the human eye (Figure 2e) compared to attacks enhanced on other channels (Figure 2i, Figure 2j). Secondly, different images exhibit different visual insensitivity to the specific trigger. For example, we can observe that the MultiBpp attack can still maintain more invisibility to the human visual system by poisoning images in Figure 2e compared to images in other images (e.g., image in Figure 5b). However, the image in Figure 2b is more visually insensitive for Blended-C compared to the images in Figure 2e. Therefore, the stealthiness of the trigger can be effectively preserved by carefully selecting appropriate samples based on the characteristics of the trigger pattern.

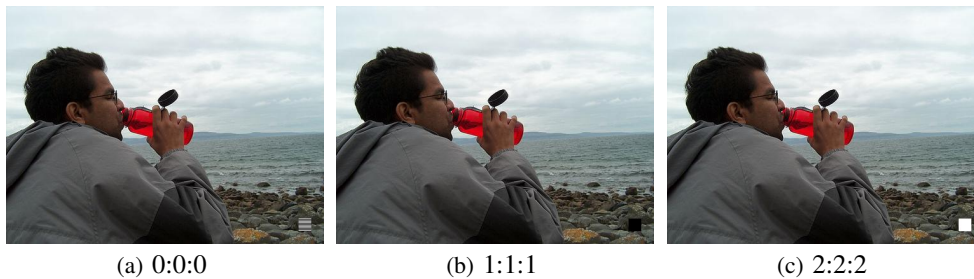


Figure 3: Visualizations of different trigger patterns in Badnets attacks. Specifically, we use  $\{0,1,2,3\}$  to represent  $\{\text{black and white striped, all-black, all-white, vanilla}\}$  triggers. Furthermore,  $N_R : N_G : N_B$  represent the distinct trigger pattern applied in  $R : G : B$  channels.

### 306 The effect of triggers in Badnets Attacks

Table 2: Performance of Badnets attacks upon CIFAR-10 with 1% samples poisoned.

Trigger			Poisoning Rate $\alpha = 1\%$				Poisoning Rate $\alpha = 2.5\%$			
Pattern			Random		Res- $x^2$		Random		Res- $x^2$	
Type	no.	Method	ASR	BA	ASR	BA	ASR	BA	ASR	BA
RGB	a	0:0:0	<b>41.99</b>	94.16	<b>90.74</b>	94.11	<b>78.29</b>	94.68	<b>92.97</b>	<b>94.58</b>
	b	1:1:1	12.13	94.23	70.00	<b>94.13</b>	34.09	94.88	74.63	94.36
	c	2:2:2	10.42	94.08	60.79	93.81	37.37	94.48	80.04	94.24
	d	1:1:0	37.31	<b>94.45</b>	86.15	93.90	63.62	94.92	89.52	94.51
	e	2:2:0	20.50	94.31	83.08	94.10	71.80	<b>94.97</b>	90.36	94.29
B	f	3:3:0	<b>40.92</b>	94.79	<b>74.74</b>	94.86	<b>68.05</b>	<b>94.75</b>	<b>91.23</b>	94.01
	g	3:3:1	12.15	94.05	53.42	<b>94.97</b>	26.47	94.39	70.80	94.42
	h	3:3:2	28.80	<b>94.96</b>	60.85	94.63	49.75	94.62	68.49	<b>94.52</b>

**The effect of triggers in Badnets attacks** As depicted in Table 2 a,f, Badnets attained a notably higher ASR when employing a black-and-white trigger compared to monochromatic triggers (all-black, all-white). Concurrently, the distinctive nature of the black-and-white trigger poses a greater challenge in identifying appropriate images for trigger concealment with Component B. According to the results between b,c and d,e, incorporating more pronounced trigger features exclusively within the blue channel also elevates the ASR in Badnets attacks. **Consequently, integrating robust features solely into the blue channel, which exhibits lower sensitivity to human perception, offers an effective means of harmonizing the benefits of both strategies.**

## G Details of Experiment Setting

Table 3: Hyperparameters and settings used in {CIFAR-10 (Krizhevsky et al. [2009]), CIFAR-100 (Krizhevsky et al. [2009]), Tiny-ImageNet (Russakovsky et al. [2015])}. sign denotes that the transformations/ augmentations are randomized.

Dataset	CIFAR-10	CIFAR-100	Tiny-ImageNet
# of Classes	10	100	200
Input Size	(3, 32, 32)	(3, 32, 32)	(3, 64, 64)
# of Images	50000	50000	100000
Target Class	0 (Airplane)	0 (Apple)	0 (Goldfish)
Epochs	200	200	400
Optimizer	SGD (Stich et al. [2018])	SGD (Stich et al. [2018])	SGD (Stich et al. [2018])
Augmentation	[Crop, H-Filp]	[Crop, Rotation]	[Crop, Rotation, H-Filp]
Model	Resnet18	Resnet18	Resnet18

**Dataset and Model** We conduct experiments on three benchmark datasets, including CIFAR-10, CIFAR-100, and Tiny-ImageNet. ResNet18 is the default model used to train the poisoned dataset. Among all datasets, the first class ( $y = 0$ ) is designated as the target class. The target class of each dataset is fixed across all the attacks adopting it. Standard augmentations are adopted on each dataset to increase the model performance following existing training pipelines (He et al. [2016a], Tan and Le [2019]). Details of the dataset can be seen in Table 3.

**Attack Setup** Three types of backdoor attacks {Badnets, Blended, BppAttack} are used as baselines to demonstrate the generalization ability of our components in {local high-intensity poisoning attacks, global medium-intensity poisoning attacks, global low-intensity poisoning attacks}.

Firstly, for BadNets attacks, a  $3 \times 3$  random noise checkerboard pattern is utilized as the trigger in CIFAR-10 and CIFAR-100. For Tiny-ImageNet, a  $9 \times 9$  is utilized as the trigger in BadNets attacks.  $\{0,1,2,3\}$  represents the distinct {black and white striped, all-black, all-white, vanilla} trigger pattern and  $N_R : N_G : N_B$  represent the distinct trigger pattern applied in  $R : G : B$  channels, as depicted in Figure 3. The origin Badnets attack can be seen as attacks with whole-black triggers (1 : 1 : 1). The Badnets trigger optimized by Component C can be represented as (1 : 1 : 0). The experiments about Component A follow the same setting as the origin work Gao et al. [2023], in which the Badnets trigger can be seen as (0 : 0 : 0).

Secondly, for Blended attacks, a Hello-Kitty image is selected as the trigger and blended with the original images.  $N_R : N_G : N_B$  represents the distinct trigger intensity applied in  $R : G : B$  channels. The default of Blended attacks can be seen as attacks with a transparency parameter of  $0.2 : 0.2 : 0.2$ . The Blended trigger optimized by Component C can be represented as  $(0.2 : 0.1 : 0.3)$ .

Furthermore, in MultiBpp attacks, the ratio  $(N_p^R : N_p^G : N_p^B)$  denotes the specific quantization configuration for poisoning intensity across the RGB channels. Notably, the default bit depth employed by BppAttack in the original study is set at 5, which, in the context of this paper, corresponds to a quantization ratio of  $32 : 32 : 32$ . Consequently, the "Base" scenario in our analysis refers to a quantization attack executed with the  $32 : 32 : 32$  ratio, excluding any training control mechanisms or label flipping operations inherent to the BppAttack methodology.

## H Extended Ablation Study

Features in backdoor attacks can be classified into {trigger feature, target-label feature, feature in non-target classes}. Given the trigger feature adjustable, the proposed components {Component A, Component B, Component C} mainly explore the potential of the inner relation between {trigger feature, feature in non-target classes}, {trigger feature, target-label feature} and {trigger feature, trigger feature}. Overall visualization can be seen in Figure 4.

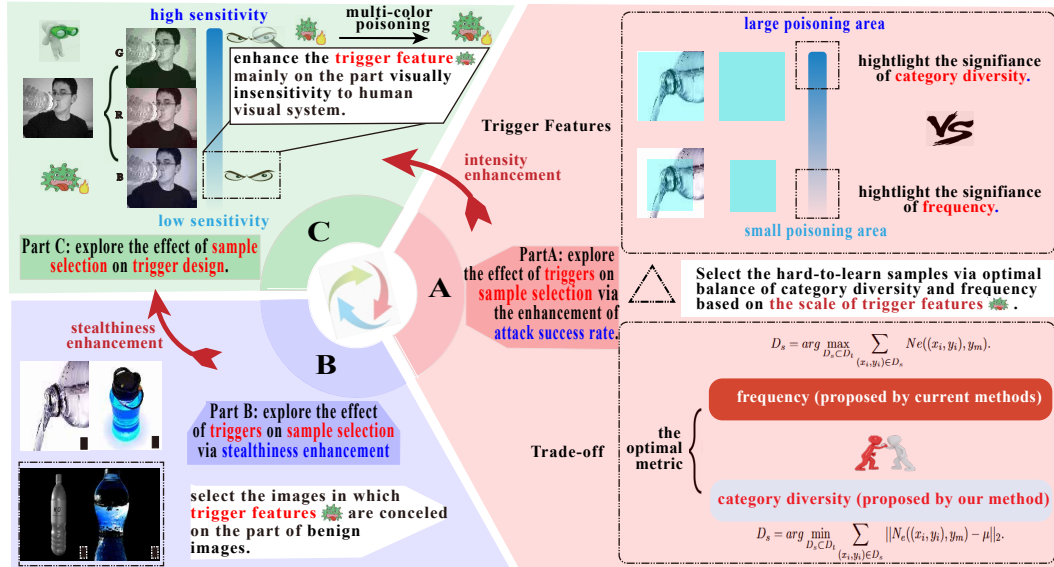


Figure 4: Overall visualization of our proposed components.

### H.1 Effect of Target Label

Target Label : 0			Target Label : 10			Target Label : 20		
Selection	ASR	BA	Selection	ASR	BA	Selection	ASR	BA
Forget	59.39	78.21	Forget	85.4	78.11	Forget	59.4	<b>78.8</b>
Res-x	<b>80.48</b>	<b>78.25</b>	Res-x	<b>91.68</b>	<b>78.12</b>	Res-x	<b>73.48</b>	78.5
Target Label : 30			Target Label : 40			Target Label : 50		
Selection	ASR	BA	Selection	ASR	BA	Selection	ASR	BA
Forget	72.94	78.31	Forget	93.23	<b>78.74</b>	Forget	82.3	<b>78.69</b>
Res-x	<b>75.83</b>	<b>78.41</b>	Res-x	<b>96.28</b>	78.27	Res-x	<b>89.46</b>	78.45
Target Label : 60			Target Label : 70			Target Label : 80		
Selection	ASR	BA	Selection	ASR	BA	Selection	ASR	BA
Forgetting Event	38.78	78.5	Forgetting Event	<b>81.96</b>	<b>78.56</b>	Forgetting Event	88.46	<b>78.61</b>
Res-x	<b>46.57</b>	<b>78.68</b>	Res-x	79.51	78.55	Res-x	<b>89.1</b>	78.32

Table 4: Performance of Badnets-C with different target labels in CIFAR-100.

**Result Analysis** To explore the effectiveness of the proposed strategy (e.g., **Res- $x$** ) on different target labels, we select labels ( $y \in \{0, 10, 20, \dots, 80\}$ ) from CIFAR-100 and 20% of the samples from the target class (representing 0.2% of the total samples) are poisoned for Badnets-C.

As depicted in Table 4, Component A exhibits a higher ASR compared to the existing state-of-the-art metric, Forgetting Event (Forget), across an overwhelming majority of experimental conditions. Notably, a substantial variation in the efficacy of backdoor attacks and corresponding defensive filtering mechanisms is contingent upon the specific target class under consideration. To illustrate, the attack success rate of the Badnets model exhibits a stark contrast, registering at 46.57% when the target class is 60, yet surging to an impressive 96.28% when the target class is 40. Furthermore, the application of our method yields a notable enhancement of 21 percentage points in performance when the target class is 0, conversely experiencing a marginal decline of 3 percentage points when the target class is 70. **Therefore, Component A exhibits the widespread applicability and robust superiority upon ASR enhancement across diverse target labels.**

## H.2 Category Similarity

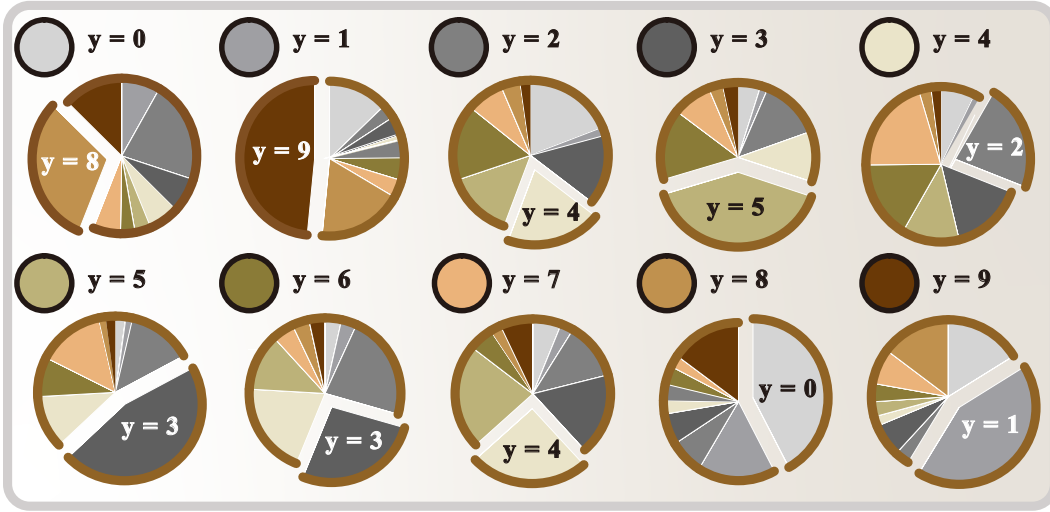


Figure 5: Category stability in CIFAR-10. We systematically arrange the proportions of misclassified categories across various data categories  $y$ , emphasizing the most prevalent category through white text highlighting. Above each visualization, the correct category corresponding to the pie chart, as well as its representative color in the context of other pie charts, is distinctly labeled.

In CIFAR-10, the correspondence between  $y$  and the true labels is  $\{0:\text{airplane}, 1:\text{automobile}, 2:\text{bird}, 3:\text{cat}, 4:\text{deer}, 5:\text{dog}, 6:\text{frog}, 7:\text{horse}, 8:\text{ship}, 9:\text{truck}\}$ . Samples of class A but frequently misclassified as class B suggest a high level of similarity between A and B. As illustrated in Figure 5, significant variations exist in the similarity among different categories. For instance, the proportion of trucks ( $y=9$ ) is substantially larger than that of birds ( $y=2$ ). Therefore, automobiles ( $y=1$ ) exhibit a much higher similarity to trucks than to birds in the pie chart representing automobiles ( $y=1$ ). Furthermore, the similarity pattern displays symmetry. For the set  $y=\{0, 1, 2, 3, 4, 5, 8, 9\}$ , the class with the highest proportion in its corresponding pie chart also dominates the pie chart of the corresponding class. Although  $y=\{6, 7\}$  deviate from this trend, they still occupy the second highest proportion in the corresponding pie charts  $y=\{3, 4\}$ .

In this paper, drawing inspiration from entropy theory, we postulate that samples belonging to class  $y$  (e.g., automobile) but frequently misclassified into a dissimilar class (e.g., bird) is more challenging to be learned by models and potentially more valuable than samples misclassified into a similar class (e.g., truck). For example, as depicted in Figure 7, samples belonging to  $\{0: \text{airplane}\}$  are often misclassified as  $\{8: \text{ship}\}$  (31.10%) rather than  $\{8: \text{frog}\}$  (3.13%) because of the distinct similarity. We hypothesize that samples misclassified as  $\{8: \text{frog}\}$  may be more informative and therefore should be considered in sample selection. Component B balances Forgetting Events and

381 Category Diversity in sample selection. **Poisoning selected samples encourages the model to adopt**  
382 **shortcuts, facilitating the learning of the trigger feature.**

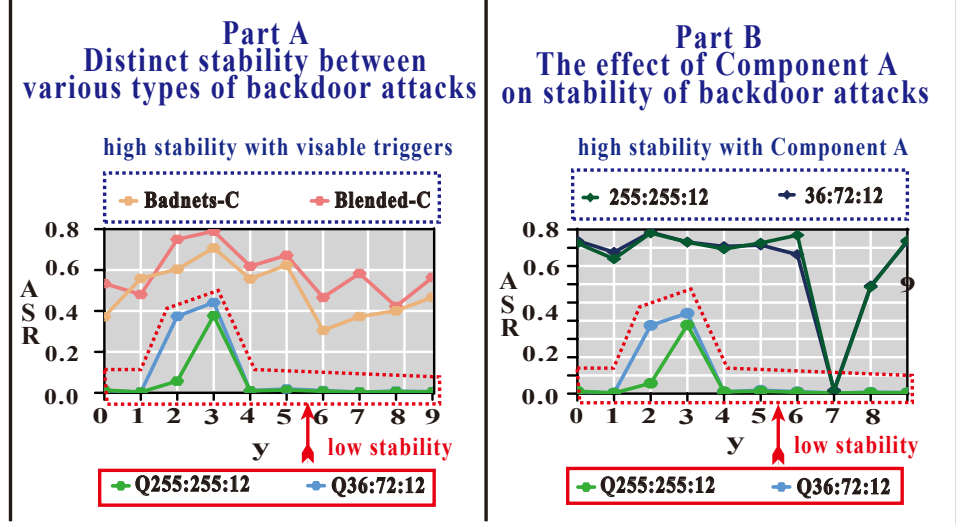


Figure 6: Inference of category similarity on backdoor attacks in CIFAR-10. We assessed the attack success rates of a conventional method characterized by prominent trigger traits and an inconspicuous method with subtle trigger characteristics to elucidate the influence of target label on attack stability.

383 What is more, it is important to note that the similarity metrics depicted in Figure 5 represent the  
384 relative proportion of similarity to other classes and do not accurately reflect the absolute magnitude  
385 of similarity between two classes. As shown in Figure 6, in the context of randomly selected data,  
386 traditional methods with pronounced trigger feature exhibit superior attack performance (maxima) at  
387  $y = \{3, 5\}$  (cat, dog). Conversely, they demonstrate inferior attack performance at  $y = 0, 6$  (airplanes,  
388 frog). For quantization methods with a subtle trigger feature, the attack is entirely ineffective in  
389 scenarios other than  $y = \{2, 3\}$ . Our analysis posits that this phenomenon arises because  $\{3, 5\}$   
390 ( $\{\text{cat, dog}\}$ ) are both small-to-medium-sized animals, sharing high similarity in color and body  
391 shape, which collectively drives the model to prioritize the learning of these two classes. In contrast,  
392 frogs exhibit lower similarity compared to other animal categories and fail to establish a synergistic  
393 relationship with them. Similarly, when juxtaposed with other modes of transportation (automobile,  
394 ship, truck), airplanes possess fewer common attributes with frogs, resulting in diminished model  
395 attention toward this category. Consequently, the trigger feature embedded in classes with reduced  
396 model attention receives less focus, ultimately undermining the efficacy of backdoor attacks.

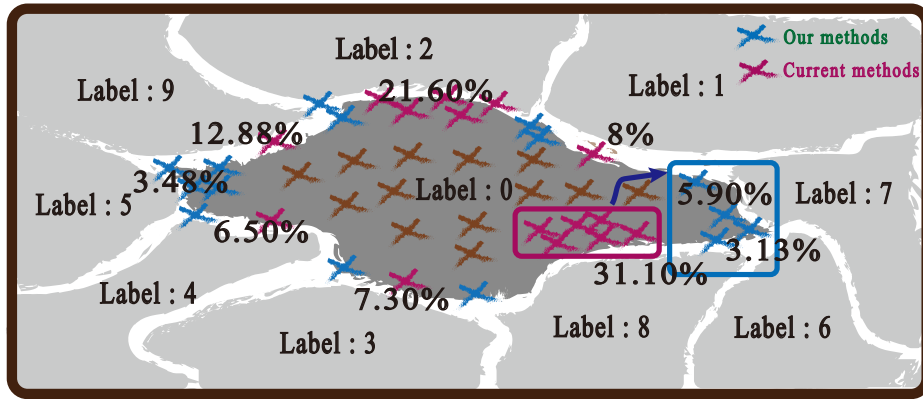


Figure 7: Visualization of misclassification results of label 0 in CIFAR-10. We use the edge lengths and associated numerical values to indicate the proportion of category information in Forgetting Event according to the pretraining stage.



Based on this analysis, it can be inferred that in scenarios involving randomly selected poisoned data, the efficacy of backdoor attacks exhibits substantial fluctuations tied to class characteristics, reflecting inadequate stability. Our findings reveal that upon adopting Component A, the merits of backdoor attacks manifest in the following ways: (1) Component A can enhance the ASR of backdoor attacks. For instance, with  $y = 0$ , the ASR of BadNets attacks with Component A gets a 40 percentage point enhancement. (2) Component A can ensure the stability of backdoor attacks. The variability in attack success rates across different classes is notably mitigated when our methodology is implemented. Notably, for quantization attacks characterized by a feeble trigger feature, Component A leads the models to circumvent class-specific constraints and detect the embedded backdoor patterns. However, while an effective trigger strategy can attenuate, but not entirely eradicate, the impact of class-related factors, quantization attacks remain ineffective in the context of  $y = 7$  (horse).

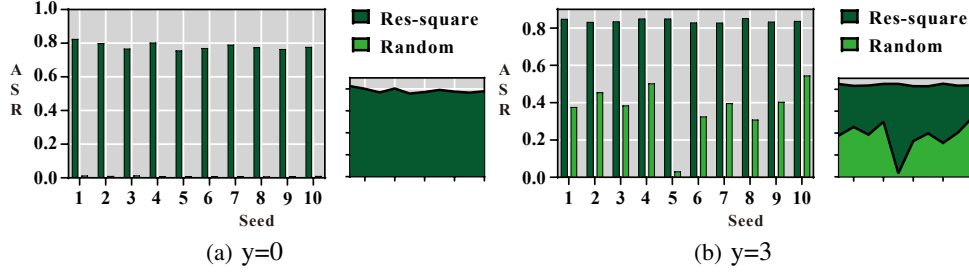


Figure 8: Stability of Badnets attacks by selecting samples using **Res- $x$** .

Figure 8 delineates the failure probabilities associated with backdoor attacks aimed at the aircraft class ( $y=0$ ) and the cat class ( $y=3$ ), alongside the attack efficacy following the implementation of our devised poisoning data selection methodology, referred to as **Res- $x^2$** . The term "seed" denotes the random seed employed to regulate and replicate stochastic processes. In experiments where poisoning data is chosen arbitrarily, the model consistently fails to acquire quantized backdoor patterns across all ten iterations within the attack configuration targeting the aircraft class. Conversely, in the scenario focusing on the cat class, the model achieves a 90% likelihood of acquiring a quantized backdoor feature, with a singular instance of failure observed under the condition of seed=5.

The aforementioned observations underscore that the model's capacity to assimilate backdoor feature is contingent upon random variables, exhibiting a strong correlation with class-specific attributes. A judicious poisoning data selection strategy can markedly bolster the robustness of the attack. Upon integrating our proposed **Res- $x^2$**  poisoning data selection strategy, the model demonstrates a flawless 100% success rate in learning backdoor feature over twenty iterations, thereby significantly attenuating the detrimental influence exerted by the choice of target class on backdoor attacks.

### H.3 Applying our methods to poisoned-label backdoor attacks

In this section, we examine the applicability of these strategies to enhance poisoned-label backdoor attacks. In the poisoned-label scenario, the selection of poisoned samples is conducted across the entire training dataset rather than being confined to the target class.  $\{0.05\%, 0.1\%, 0.15\%, 0.2\%\}$  of the total samples are poisoned for Badnets. We evaluate our plug-in methods (**Res- $x$** ) against the standard version with random selection (dubbed 'Random').

As illustrated in Table 5, the effect of selecting samples based on the Forgetting Event metric on attacks exhibits a similar performance as our methods. We speculate the reasons is that the sample selection used in a clean-label setting aims to identify the most challenging samples of the target category, thereby incentivizing the model to acquire and leverage backdoor feature. However, when the scope of poisoned data selection encompasses the entire dataset, data originating from categories distinct from the current target category proves to be more "challenging to train" than data inherent to the target category. Consequently, these data points are more likely to induce the model to concentrate on the backdoor feature. For example, we delve into the intricacies of our category information calculation methodology. This process involves, for a given target category to elucidate the underlying mechanisms contributing to this phenomenon, computing the distribution of misclassification information, which encapsulates the extent to which data from that category

Table 5: Performance of poison-label attacks in CIFAR-100 with different poisoning rates.

<i>Poisoning Rate : 0.05%</i>				<i>Poisoning Rate : 0.1%</i>			
Attack	Selection	ASR	BA	Attack	Selection	ASR	BA
Badnets	Forget	<b>7.00</b>	78.26	Badnets	Forget	<b>53.98</b>	78.51
	<b>Res-<math>x</math></b>	<b>27.55</b>	<b>78.59</b>		<b>Res-<math>x</math></b>	<b>54.53</b>	<b>78.55</b>
Blend	Forget	<b>64.07</b>	78.31	Blend	Forget	<b>73.29</b>	<b>78.70</b>
	<b>Res-<math>x</math></b>	62.66	<b>78.74</b>		<b>Res-<math>x</math></b>	71.33	78.47
<i>Poisoning Rate : 0.15%</i>				<i>Poisoning Rate : 0.2%</i>			
Attack	Selection	ASR	BA	Attack	Selection	ASR	BA
Badnets	Forget	59.75	78.54	Badnets	Forget	74.3	<b>78.85</b>
	<b>Res-<math>x</math></b>	<b>68.43</b>	<b>78.95</b>		<b>Res-<math>x</math></b>	<b>79.77</b>	78.77
Blend	Forget	<b>78.22</b>	<b>78.27</b>	Blend	Forget	83.04	<b>78.58</b>
	<b>Res-<math>x</math></b>	77.52	78.04		<b>Res-<math>x</math></b>	<b>84.74</b>	78.47

are erroneously classified into alternative categories. Component A facilitates the sample selection of attacks where the target category aligns with the image’s inherent category. Conversely, this methodology is less conducive to attacks where the target category diverges from the image’s category, thereby underscoring the inherent limitations of our approach in poison-label attacks. Furthermore, we observe that ASRs of our methods exhibit consistently superior performance against Forgetting Event for BadNets attacks, in contrast to Blended attacks, under both poison-label and clean-label settings. We speculate that it is mostly because of the trigger-related feature (e.g., small poisoning area) of badnets. We will further explore its mechanism in our future work.

#### H.4 The effect of Component A in Tiny-Imagenet

Table 6: Current methods on Tiny-ImageNet.

Method	Metric	Badnets-C	Blended-C
Vanilla	BA	57.50%	57.27%
	ASR	17.06%	27.71%
Loss Value	BA	57.17%	57.49%
	ASR	32.22%	37.63%
Gradient Norm	BA	57.69%	57.82%
	ASR	31.74%	38.74%
Forgetting Event	BA	57.60%	57.48%
	ASR	<b>32.29%</b>	<b>40.59%</b>

Table 7: Our methods on Tiny-ImageNet.

Method	Metric	Badnets-C	Blended-C
res-log	BA	57.03%	57.24%
	ASR	34.46%	42.02%
res-linear	BA	57.17%	57.16%
	ASR	32.22%	41.48%
res-square	BA	58.01%	57.02%
	ASR	<b>38.96%</b>	<b>43.93%</b>
res-exp	BA	57.60%	57.58%
	ASR	32.29%	38.31%

**Analysis on Tiny-imagenet** We conduct experiments on Tiny-Imagenet, which is a simplified version of Large Scale Visual Recognition Challenge 2016 Russakovsky et al. [2015], with ResNet-18 (He et al. [2016b]). We compare our plug-in methods against the standard version with random selection (dubbed ‘vanilla’) and existing sample selection strategies based on current metrics (such as forgetting events, gradient norm, and loss value). Across all these attacks upon Tiny-imagenet, 50% of the samples from the target class (representing 0.25% of the total samples) are poisoned, with the first class designated as the target class. Results can be seen in Tables 6&7.

As depicted in Tables 6 and 7, Badnets-C achieves a 38.96% ASR under our proposed Res-square strategy, which is 6.67% higher than the current optimal traditional filtering metric (Forgetting Event). Blended-C achieves an attack success rate of 43.93% under our proposed Res-square strategy, which is 3.34% higher than the current optimal traditional filtering metric (Forgetting Event). Furthermore, we learn that Badnets-C reaches optimal attack effectiveness when adopting the more aggressive Res-square strategy on Tiny-ImageNet instead of the optimal strategy (Res-log) when trained on CIFAR10. This indicates that in the Tiny-Imagenet dataset with more categories (200), **Category Diversity should be highlighted when searching for the appropriate combination of Forgetting Event and Category Diversity in Component A.**

## 464 I Stealthiness of our components on multiple attacks

465 In this section, high-quality images of the same category in ImageNet are used to facilitate the comparison between the visibility of various methods.



Figure 9: Images poisoned by Blended attacks with different GMSD values.

466 **The effect of GMSD values in stealthiness enhancement of Blended attacks** As depicted in  
 467 Figure 9, poisoned images with lower GMSD values exhibit superiority in the stealthiness of triggers  
 468 for blended attacks. Component B with GMSD tends to find samples with complex backgrounds  
 469 where the visual sensitivity to Blended triggers will significantly weaken ( $\text{GMSD} \in [0.027, 0.080]$ ). In  
 470 contrast, the Hello Kitty triggers are easy to find when poisoned in images with a simple background  
 471 (specifically, all-white patch) where  $\text{GMSD} \in [0.471, 0.500]$ . Therefore, **Component B with GMSD**  
 472 **can significantly enhance the stealthiness of Blended attacks.**



Figure 10: Images poisoned by MultiBpp-B attacks with different GMSD values.

473



**The effect of GMSD values in stealthiness enhancement of MultiBpp attacks** As depicted in Figure 10, MultiBpp attacks exhibit satisfactory performance in Stealthiness even in the images with the lowest GMSD. Component B with GMSD tends to find samples with complex colors where the visual sensitivity to MultiBpp triggers will significantly weaken ( $\text{GMSD} \in [0.0274, 0.0769]$ ). In contrast, single-color dominated images where  $\text{GMSD} \in [0.3806, 0.4927]$  are selected by Component B to serve as suboptimal samples. The results of lower GMSD suggest that single-channel color variations may amplify susceptibility to MultiBpp attacks under extreme conditions. Specifically, attenuation of intensity in the green channel (a region of heightened visual sensitivity in the human visual system) and elevation in the blue channel (a region of reduced sensitivity in the human visual system) both result in lower GMSD values. Therefore, single-color dominated samples will not be selected for poisoning. In summary, **Component B with GMSD can significantly enhance the stealthiness of MultiBpp attacks and benefit the performance of Component C.**

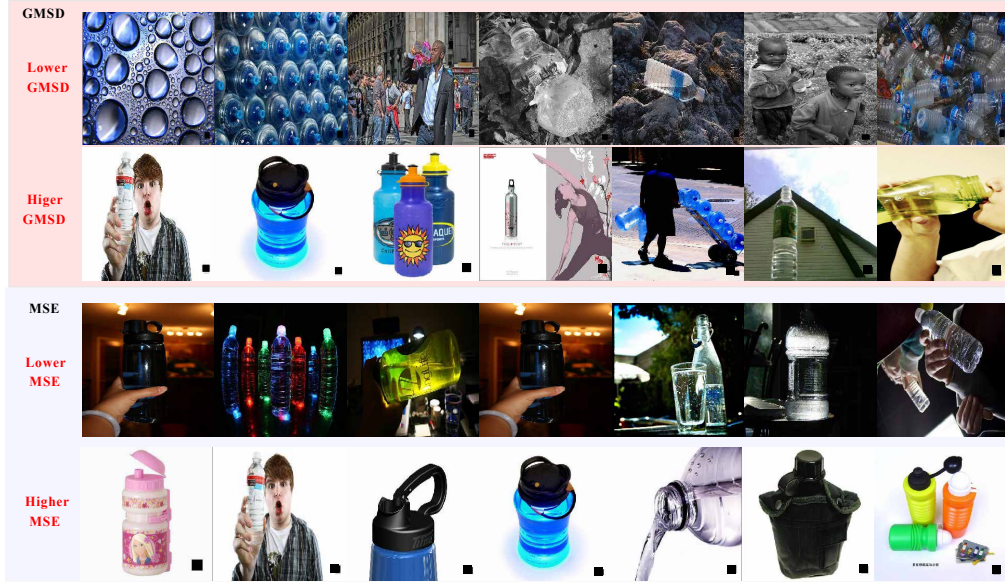


Figure 11: Images poisoned by Badnets attacks with different evaluation metrics in Component B.

**The effect of different metrics in stealthiness enhancement of Badnets attacks** As depicted in Figure 11, poisoned images with lower GMSD values and lower MSE both exhibit superiority in the stealthiness of triggers for badnets attacks. Component B with GMSD tends to find samples with complex colors where the visual sensitivity to MultiBpp triggers will significantly weaken ( $\text{GMSD} \in [0.0274, 0.0769]$ ). In contrast, single-color dominated images where  $\text{GMSD} \in [0.3806, 0.4927]$  are selected by Component B to serve as suboptimal samples. Component B with GMSD tends to find samples with complex backgrounds where the visual sensitivity to Badnets triggers will weaken. In contrast, Component B with MSE tends to find samples with patches similar to the triggers where the visual sensitivity to Badnets triggers will significantly weaken. Therefore, MSE exhibits superiority in the stealthiness enhancement of Badnets attacks compared to GMSD and will be applied in this paper. In summary, **Component B with MSE and GMSD can significantly enhance the stealthiness of Badnets attacks and benefit the performance of Component C.**

## References

- Jiawang Bai, Kuofeng Gao, Dihong Gong, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Hardly perceptible trojan attack against neural networks with bit flips. In *European Conference on Computer Vision*, pages 104–121. Springer, 2022.
- Peng Chen, Jirui Yang, Junxiong Lin, Zhihui Lu, Qiang Duan, and Hongfeng Chai. A practical clean-label backdoor attack with limited information in vertical federated learning. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 41–50. IEEE, 2023.

505 Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep  
506 learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

507 Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack  
508 framework for diffusion models. *Advances in Neural Information Processing Systems*, 36:33912–  
509 33964, 2023.

510 Yinghua Gao, Yiming Li, Linghui Zhu, Dongxian Wu, Yong Jiang, and Shu-Tao Xia. Not all samples  
511 are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition*, 139:109512,  
512 2023.

513 Yinghua Gao, Yiming Li, Xueluan Gong, Zhifeng Li, Shu-Tao Xia, and Qian Wang. Backdoor attack  
514 with sparse and invisible trigger. *IEEE Transactions on Information Forensics and Security*, 2024.

515 Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the  
516 machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

517 Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tianwei  
518 Zhang. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy (SP)*,  
519 pages 3385–3403. IEEE, 2024.

520 Jonathan Hayase and Sewoong Oh. Few-shot backdoor attacks via neural tangent kernels. *arXiv  
521 preprint arXiv:2210.05929*, 2022.

522 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image  
523 recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition  
524 (CVPR)*, June 2016a.

525 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image  
526 recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,  
527 pages 770–778, 2016b.

528 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

529 Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical society of  
530 America*, 61(1):1–11, 1971.

531 Haoyang Li, Qingqing Ye, Haibo Hu, Jin Li, Leixia Wang, Chengfang Fang, and Jie Shi. 3dfed:  
532 Adaptive and extensible framework for covert backdoor attack in federated learning. In *2023 IEEE  
533 Symposium on Security and Privacy (SP)*, pages 1893–1907. IEEE, 2023.

534 Sen Li, Junchi Ma, and Minhao Cheng. Invisible backdoor attacks on diffusion models. *arXiv  
535 preprint arXiv:2406.00816*, 2024.

536 Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible back-  
537 door attacks on deep neural networks via steganography and regularization. *IEEE Transactions on  
538 Dependable and Secure Computing*, 18(5):2088–2105, 2021. doi: 10.1109/TDSC.2020.3021407.

539 Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite backdoor attack for deep neural  
540 network by mixing existing benign features. In *Proceedings of the 2020 ACM SIGSAC conference  
541 on computer and communications security*, pages 113–131, 2020.

542 Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Backdoor attacks against dataset  
543 distillation. *arXiv preprint arXiv:2301.01197*, 2023.

544 Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assump-  
545 tion of latent separability for backdoor defenses. In *The Eleventh International Conference on  
546 Learning Representations*, 2023. URL [https://openreview.net/forum?id=\\_wSHsgrVali](https://openreview.net/forum?id=_wSHsgrVali).

547 Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang,  
548 Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition  
549 challenge. *International journal of computer vision*, 115:211–252, 2015.

550 Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory.  
551 In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, ed-  
552 itors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates,  
553 Inc., 2018. URL [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/](https://proceedings.neurips.cc/paper_files/paper/2018/file/b440509a0106086a67bc2ea9df0a1dab-Paper.pdf)  
554 [b440509a0106086a67bc2ea9df0a1dab-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/b440509a0106086a67bc2ea9df0a1dab-Paper.pdf).

555 Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks.  
556 In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International*  
557 *Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages  
558 6105–6114. PMLR, 09–15 Jun 2019. URL [https://proceedings.mlr.press/v97/tan19a.](https://proceedings.mlr.press/v97/tan19a.html)  
559 [html](https://proceedings.mlr.press/v97/tan19a.html).

560 Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Invisible  
561 black-box backdoor attack against deep cross-modal hashing retrieval. *ACM Transactions on*  
562 *Information Systems*, 42(4):1–27, 2024.

563 Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against  
564 deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings*  
565 *of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15074–15084,  
566 2022.

567 Emily Wenger, Roma Bhattacharjee, Arjun Nitin Bhagoji, Josephine Passananti, Emilio Andere,  
568 Heather Zheng, and Ben Zhao. Finding naturally occurring physical backdoors in image datasets.  
569 *Advances in Neural Information Processing Systems*, 35:22103–22116, 2022.

570 Yutong Wu, Xingshuo Han, Han Qiu, and Tianwei Zhang. Computation and data efficient backdoor  
571 attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages  
572 4805–4814, 2023.

573 Yuan Xun, Xiaojun Jia, Jindong Gu, Xinwei Liu, Qing Guo, and Xiaochun Cao. Minimalism is king!  
574 high-frequency energy-based screening for data-efficient backdoor attacks. *IEEE Transactions on*  
575 *Information Forensics and Security*, 2024.

576 Shuai Zhao, Luu Anh Tuan, Jie Fu, Jinming Wen, and Weiqi Luo. Exploring clean label backdoor  
577 attacks and defense in language models. *IEEE/ACM Transactions on Audio, Speech, and Language*  
578 *Processing*, 2024.

579 Zihao Zhu, Mingda Zhang, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Boosting backdoor  
580 attack with a learnable poisoning sample selection strategy. *arXiv preprint arXiv:2307.07328*,  
581 2023.